DOCUMENT RESUME

ED 252 574 TM 850 050

AUTHOR Friedman, Larry P.
TITLE Trivariate Scatter Plots.
PUB DATE Aug 84
NOTE 35p.; Paper presented at the Annual Meeting of the American Psychological Association (92nd, Toronto, Ontario, August 24-28, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Graphics; Computer Software; Correlation; *Data Analysis; Factor Structure; Monte Carlo Methods; *Multivariate Analysis; *Statistical Distributions
IDENTIFIERS Data Display; Scatter Plots; TAU Equivalence; THREEDE Computer Program

ABSTRACT

Few methods have been tried and used to graphically represent more than two variables. This poster session showed a new method for representing three continuous variables on a single scatterplot using the THREEDE computer program. Two variables are represented as a normal bivariate distribution. The third variable is represented by a symbol, e.g. a circle or square, surrounding the bivariate point. The symbol size shows the relative size of the third variable. Additional symbols or different colors can be used. Methods and examples for three statistical areas are presented: (1) correlational data between three variables; (2) a TAU squared design (between-level differences of three dependent variables for an independent variable); and (3) factor loadings. (Author/BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED252574

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Trivariate Scatter Plots

Larry P. Friedman

California State University Northridge

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. P. Friedman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 050

Paper presented at the 92nd annual American Psychological Association convention, August, 1984, in Toronto, Canada.

Running head: TRIVARIATE SCATTER PLOTS

Abstract

Few methods have been tried and used to graphically represent more than two variables. This poster session shows a new method that is used to represent three continuous variables on a single scatter plot. The variables are displayed in a way analogus to a bivariate scatter plot. One group presentation shows how the basic scatter plot looks when the three variables are related to differening degrees. Two and three groups can be contrasted using different colors or symbols, e.g. a circle, a pentagon, and/or a square. The relations among columns of factor patterns are also investigated with this method.

Trivariate Scatter Plots

Two dimensional scatter plots are very common descriptors in psychology today. Few methods have been tried and used to represent more than two variables graphically. Ripley (1981) uses multivariate data to create contours and grids. Other methods have also been used to show multivariate data. In the 1960's, John Tukey and Mary Ann Fisnerkeiler (Kolata, 1982) created a computer graphics system that displays multivariate data, three variables at a time. Depth is shown by color. A computer terminal with graphics capabilities must be used with this. Another way of displaying multivariate data is with Chernoff faces (Bruckner, 1978). Each dimension is represented by one of the eighteen facial features. Bruckner also talks about Metroglyphs and Andrews sine curves as other methods of display.

In the current study, three dimensional plotting is to be investigated with the representations I have developed. A computer program, THREEDE, was created to make the plots. This plotting shows the relationship of three continuous variables in a way analogus to a bivariate scatterplot. Two of the variables are represented on the plot in the normal way as a bivariate distribution. The third variable is represented by a symbol e.g. a circle, surrounding the bivariate point. The size of the symbol shows the relative size of the third variable.

This poster session presents a number of statistical areas in which such representation could be used. For example, one could use this method in representing correlational data between three variables. One could also use this procedure in showing "between-level-differences" or three dependent variables for an independent variable, as in a TAU squared design. It could also be used to show how a group would look after treatment in which there is an increase on one variable and a decrease on another. Another application could be in the representation of three columns or factor loadings before and after rotation.

Methods: One Group

This section discusses how the data is created and how one group with three dependent variables is displayed using this plotting technique. The correlations that are discussed throughout the paper are the correlations between variable 1 and variable 2, variable 1 and variable 3, and variable 2 and variable 3. Variable 1 is located on the X axis. Variable 2 is on the Y axis. Variable 3 is on the Z axis which is perpendicular to the paper, or "coming out at you".

Monte-Carlo techniques and principal component analysis techniques were used to find and create the data to be plotted. THREEDE, calls on the IMSL (International Mathematics and Science Library) library program to generate the original random data. Once obtained, the data is then transformed into data with specified correlations through

principal component computations.

A Monte-Carlo computer technique was used because it readily provides data with specific correlational relationships. The technique, which uses a random sample from a normal distribution, also avoids biases in the data.

The data that the IMSL library program creates is entered in the program THREEDE in three vectors of user specified length. The vectors have an approximately normal shape. Each vector contains the created scores for all subjects on one variable. The three vectors are slightly correlated with each other, because the IMSL library program sends back what is considered a sample population. Therefore, to be orthogonal, the data must be transformed. The data can then be transformed a second time to have the precise specified correlations.

The transformation procedure that changes the sample data into orthogonal data uses principle component analysis and is done by (1) taking the sample data (matrix X), (2) finding its eigenvalues (matrix D_1^2), (3) finding its eigenvectors (matrix E_1), and (4) deriving its orthogonal component scores (matrix G). The principal component equation for this is:

$$G = X * (D_1 * E_1')^{-1}.$$

Matrix G becomes the new orthogonal data matrix. The eigenvalues for data matrix X were obtained through a direct method. The formula solves a third degree polynomial. It

will only work if the roots are real numbers. In this case, they are always real. The original formula was a generalized formula which worked with positive definite matrices. The formula was modified to work only with correlation matrices (i.e. 1's in the diagonal). The algorithm is given in Appendix A. The eigenvectors were found through Gaussian elimination and then normed by column. The algorithm for this is presented in Appendix B.

Given the eigenvalues and eigenvectors, one can obtain a N x 3 matrix G of data points whose columns are orthogonal to each other. In addition, since their means are zero, the three vectors are uncorrelated. The matrix G may now be transformed to have the specified correlations. The method of triangular decomposition was used to obtain the rotation matrices required to transform the vectors so they had the selected correlations. The algorithm that produces the transformation matrix is provided in appendix C.

Four degrees of association were selected to use in representing the correlations because of the linear increment between values of the variances. The shared variances are 0.0, .30, .50, and .70 and the correlations are -.837, -.707, -.548, 0.0, .548, .707, and .837. The dependency between the variables were systematically varied to achieve all possible orders.

For these seven numbers there are 343 possible orders of their value taken three at a time. Of these, 175 yielded

acceptable multiple correlation combinations while 168 yielded multiple correlations greater than 1.00. A multiple correlation greater than 1.00 indicates an impossible combination of correlations and hence there combination had to be rejected. That such relations are unacceptable can be illustrate by considering three variables, of which two are positively correlated to each other at .90. If one of these two is negatively correlated to a third, the second must also be negatively correlated to the third within some variance. Or, if one of these two is positively correlated to a third, the second must also be positively correlated to the third within some variance. There are particular limits to what the correlations between three variables can be before they conflict with the logical and mathematical relationships between variables. The formula to determine the multiple correlation between three variables is as follows (see, e.g. Guildford, 1956):

$$CCC = (r_1^2 + r_2^2 - 2 * r_1 * r_2 * r_3) / (1.0 - r_3^2).$$

If CCC is greater than 1.0, it is impossible for these three variables to be related as given by the proposed combination.

Discussion

Figure 1 makes it easy to see how these plots can be used. The correlations between the variables are all high. One can see this from the fact that as the points in variable 1 on the X axis get higher in value, the points in variable 2 on the Y axis also get higher in value. This is the usual

bivariate relationship. The value of each of the data points in variable 3 is shown by the relative size of the circle around each asterisk. In figure 1, the high correlation between variable 1 and variable 3 can be seen by looking at what happens to the circles relative to the X axis. As the values on the X axis get higher, the relative size of the circles gets larger. This is similar to the process between variable 2 and variable 3 except that it is along the Y axis. Figures 2 through 6 show more exemplary relationships. When the relationship is not very obvious, it may take longer to see exactly what is going on.

Methods: Two Groups

Differences of three variables between levels of an independent variable such as in a TAU squared design has also been investigated. For instance, how a group looks after treatment in which it generally increases on one variable and generally decreases on another has been shown. In this section, Monte-Carlo techniques were also used to create the data to plot and in addition, results from selected research studies to show its practical use are presented.

The method of finding the data for two groups is essentially the same as for one group. The difference is that the procedure described in the previous section is performed twice, once for each group. Using this procedure, the correlations within each group are equal to the requested correlations, but the overall correlations of all data points

are different.

Group differences can be added to a particular group on a particular variable. The group difference is added following a formula set up by Cohen (1969). It takes into consideration the standard deviation and the size of the effect desired. Cohen shows the calculation of the size of effect parameter, delta (d), as follows:

$$d = (\bar{X}_1 - \bar{X}_2) / SD$$

Delta can be thought of as the distance between means in terms of Z scores assuming that the variables are normally distributed. A delta of 1.0 is a Z score of 1.0. A small effect shows up as a delta of .2 or .3. A medium effect is about .5 or .6. A large effect is a delta of .8 or .9.

When the data is originally created, it is standardized and then it is normed. It therefore has a mean of zero. Its standard deviation is dependent on the number of subjects in each group. The group difference is calculated by changing the previous formula slightly as follows:

$$(\bar{X}_1 - \bar{X}_2) = d * SD$$

The user simply specifies the size of the delta increase desired. The program then calculates the standard deviation and then the group difference and adds it to the group and the dependent variable specified. The effect size for each group and variable is plotted in the lower right hand corner of the plot.

A multivariate F test is automatically performed when

there are two or more groups to be plotted. It is a one way multivariate analysis of variance. When there are only two groups, it becomes a TAU squared design. The F value with its associated degrees of freedom and the probability level are plotted on the graph in the upper righthand corner. The algorithm for the matrix E(H + E) was taken from the formula found in Morrison (1967, pp. 168-169). The algorithm to find Rao's approximation of the F statistic from this matrix was taken from Marascuilo (1983, p. 315). The probability was calculated using the IMSL subroutines. The input to the subroutine is the value of F and its associated degrees of freedom. The output is the probability value of (1.0 - p).

Discussion

From these two group plots (figures 7 through 14) one can see differences between the groups. The first four, two group plots (figures 7 through 10), show group differences that can be seen on a bivariate plot. The other four, two group plots (figures 11 through 14), show group differences including the third variable.

Methods: Factors

A third area investigated is in the representation of factor loadings. The data that were used here were from the 24 psychological tests (Holzinger & Swinford, 1938). Four factors were derived from the data using the criterior of 1.0 for the lowest eigenvalue. Six plots are required to show all possible bivariate combinations of the four factors.

Four plots are required to show all possible trivariate combinations. The four plots are shown in figures 15 through 18. For ease of seeing the factor structure using the third variable on the Z axis, the absolute values of all factor loadings for the third factor of the plot were used.

Appendix A

The following is a direct method to find three eigenvalues of a correlation matrix.

```
P1 = -3.0D0
Q1 = -(C1**2 + C2**2 + C3**2 - 3.0D0)
R1 = -(2.0D0*C1*C2*C3 + 1.0D0 - C1**2 - C2**2 - C3**2)
A1 = (3.0D0 * Q1 - P1**2) / 3.0D0
B1 = (2.0D0*P1**3 - 9.0D0*P1*Q1 + 27.0D0*R1) / 27.0D0
M1 = 2.0D0 * DSQRT(-A1 / 3.0D0)
K1 = 3.0D0 * B1 / (A1 * M1)
K2 = DACOS(K1) / 3.0D0
C11 = DCOS(K2 + 2.0D0 * 3.14159265D0 / 3.0D0)
C22 = DCOS(K2 + 4.0D0 * 3.14159265D0 / 3.0D0)
C33 = DCOS(K2)
D(1,1) = M1 * C11 + 1.0D0
D(2,2) = M1 * C22 + 1.0D0
D(3,3) = M1 * C33 + 1.0D0
```

C1, C2, and C3 are the correlations between the variables. D(1,1), D(2,2), and D(3,3) become the three eigenvalues of the correlation matrix for the data matrix X.

Appendix B

The following algorithm derives three eigenvectors, one at a time for each of the eigenvalues where L(I) is each eigenvalue.

```
DO 6210 I = 1,3
N(1,1) = 1.0D0 - L(I)
N(2,2) = 1.0D0 - L(I)
N(1,2) = C1 / N(1,1)
N(1,3) = C2 / N(1,1)
N(2,3) = C3
N(2,2) = (N(1,2) * C1) - N(2,2)
N(2,3) = (N(1,3) * C1) - N(2,3)
N(2,3) = N(2,3) / N(2,2)
N(1,3) = -(N(2,3) * N(1,2)) + N(1,3)
E(1,I) = -N(1,3)
E(2,I) = -N(2,3)
E(3,I) = 1.0D0
SSEIGEN = DSQRT(E(1,I)**2 + E(2,I)**2 + 1.0D0)
E(1,I) = E(1,I) / SSEIGEN
E(2,I) = E(2,I) / SSEIGEN
E(3,I) = E(3,I) / SSEIGEN
6210 CONTINUE
```

Where C1, C2, and C3 are the correlations.

Appendix C

The following algorithm does triangular decomposition on a 3 x 3 correlation matrix to find a transformation matrix. C1, C2, and C3 are the correlations.

```
X(1,1) = 1.0D0
X(2,2) = 1.0D0
X(3,3) = 1.0D0
X(1,2) = C1
X(2,1) = C1
X(1,3) = C2
X(3,1) = C2
X(2,3) = C3
X(3,2) = C3
DO 3010 J = 1, 3
  DO 3020 I = 1, 3
    T1(I,1) = 0.0D0
    T2(1,I) = 0.0D0
    DO 3030 II = 1, 3
      L(I,II) = 0.0D0
    3030 CONTINUE
  3020 CONTINUE
  P = 1 / X(J,J)
  P = DSQRT(P)
  DO 3040 I = 1, 3
    T1(I,1) = X(I,J) * P
    T2(1,I) = X(J,I) * P
```

```
3040 CONTINUE
DO 3050 K = 1, 3
  DO 3060 KK = 1, 3
    L(K,KK) = T1(K,1) * T2(1,KK)
  3060 CONTINUE
3050 CONTINUE
DO 3070 K = 1, 3
  DO 3080 KK = 1, 3
    X(K,KK) = X(K,KK) - L(K,KK)
  3080 CONTINUE
3070 CONTINUE
DO 3090 I = 1, 3
  A(I,J) = T1(I,1)
  B(J,I) = T2(1,I)
3090 CONTINUE
3010 CONTINUE
```

References

Bruckner, L. A. (1978). On Chernoff Faces. In Graphical Representation of Multivariate Data, Ed. Wang, P. C. C. New York: Academic Press, Inc. From the "Symposium on Graphical Representation of Multivariate Data", Naval Postgraduate School, Montery, CA., 1978.

Cohen, J. (1969). Statistical Power Analysis for the Behavioral Sciences, New York: Academic Press.

Guilford, J. P. (1956). Fundamental statistics in psychology and education, New York: Mc Graw-Hill Book Company, Inc.

Holzinger, K. J. & Swineford, F. (1939). A study in factor analysis: stability of a two factor solution. Supplimentary Educational Monograph, #48.

Kolata, G. (1982, September 3). Computer Graphics Comes to Statistics. Science, 217(4563), 919-920.

Marascuilo, L. A. & Levin, J. R. (1983). Multivariate Statistics in the Social Science: A Researcher's Guide, Monterey, California: Brooks/Cole Publishing Company.

Morrison, D. F. (1967). Multivariate Statistical Methods, New York, New York: McGraw-Hill Book Company.

Ripley, B. D. (1981). Spatial Statistics, New York: John Wiley & Sons.

Figure 1.

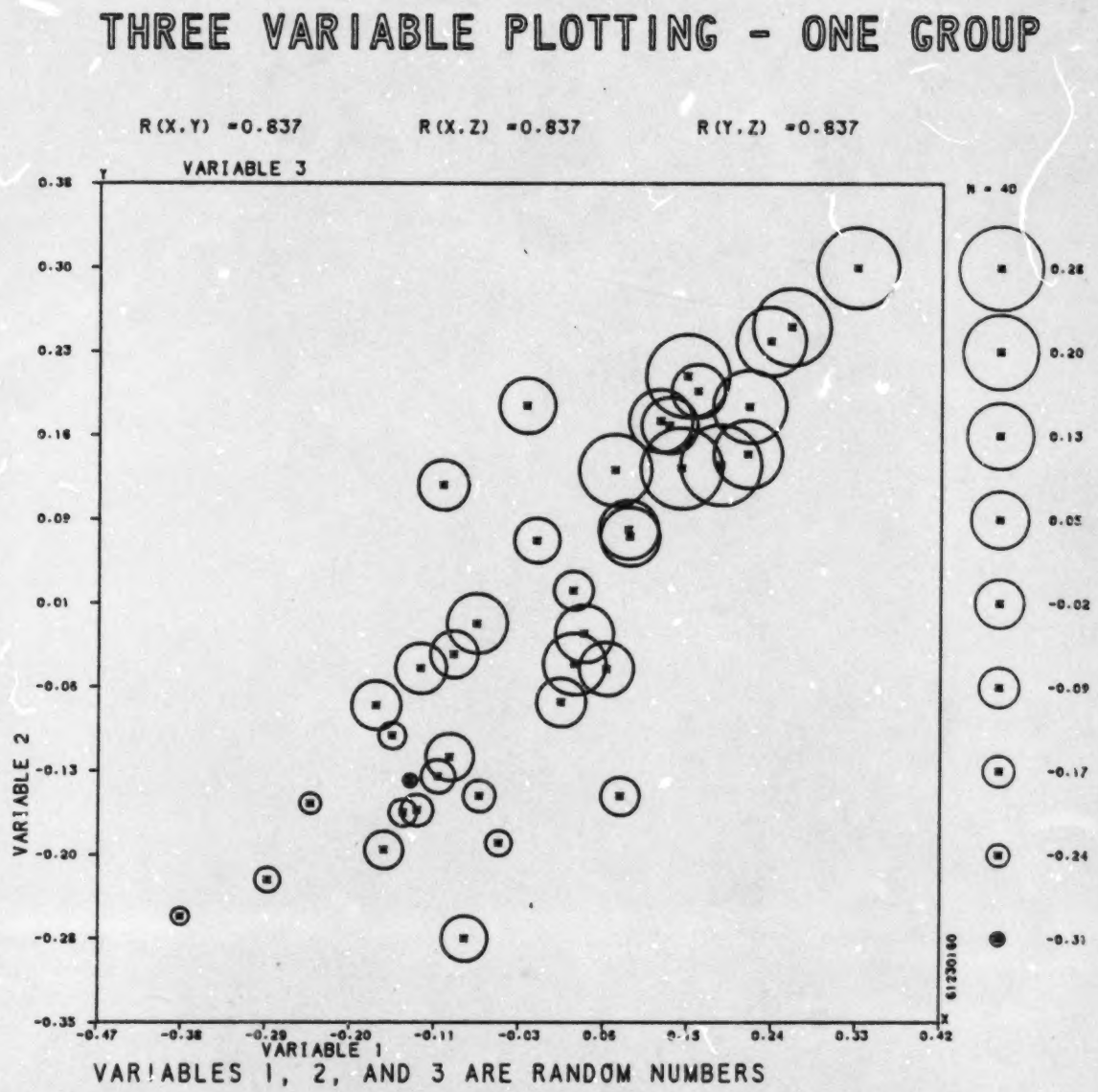

THREE VARIABLE PLOTTING - ONE GROUP
R(X,Y) =0.837
R(X,Z) =0.837
R(Y,Z) =0.837
VARIABLE 3
N = 40
VARIABLE 2
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 2.

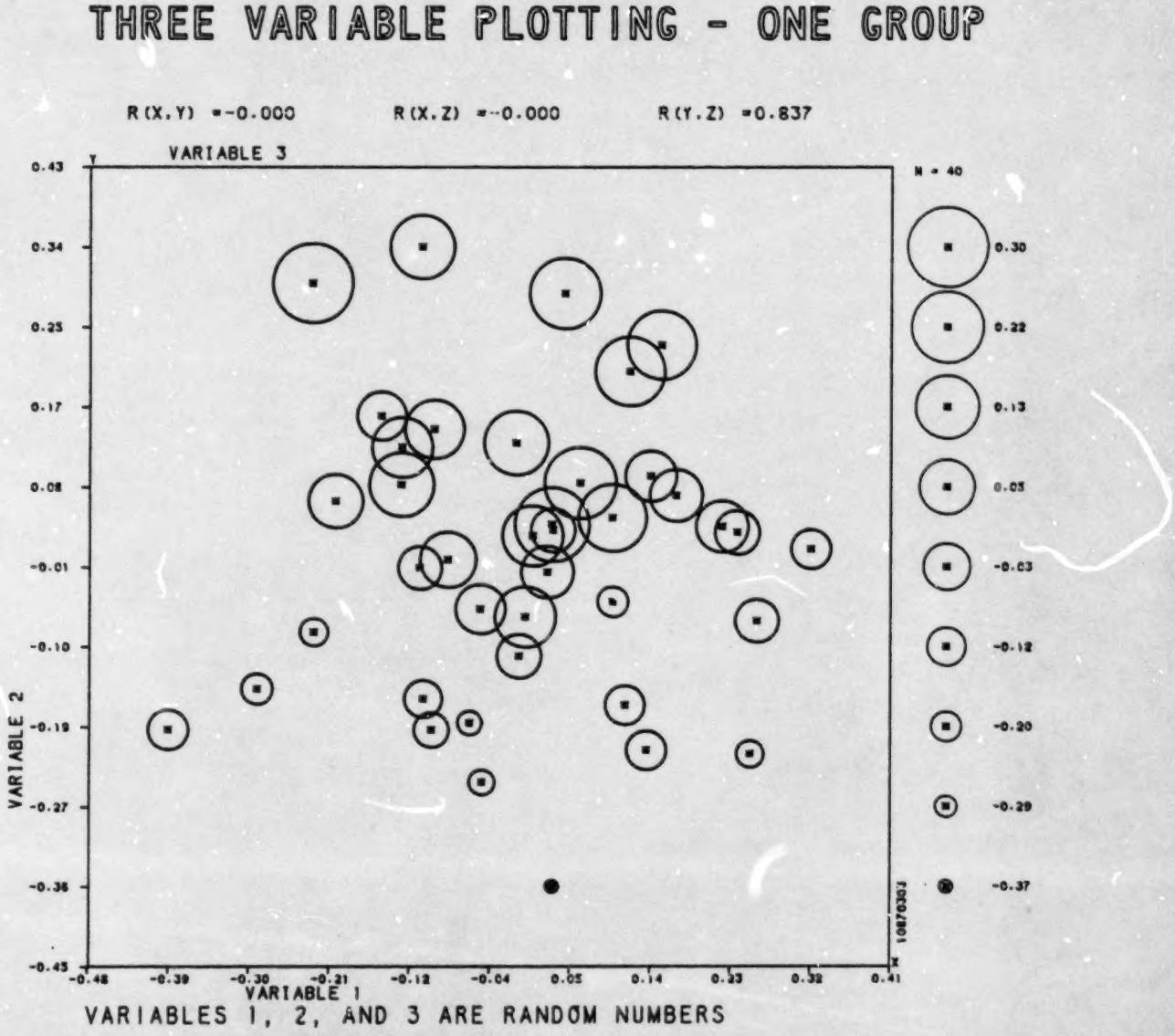

THREE VARIABLE PLOTTING - ONE GROUP
R(X,Y) =-0.000
R(X,Z) =-0.000
R(Y,Z) =0.837
VARIABLE 3
N = 40
0.43
0.34
0.25
0.17
0.08
-0.01
-0.10
-0.19
-0.27
-0.36
-0.45
0.30
0.22
0.13
0.03
-0.03
-0.12
-0.20
-0.29
-0.37
VARIABLE 2
-0.48
-0.39
-0.30
-0.21
-0.12
-0.04
0.05
0.14
0.23
0.32
0.41
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 3.

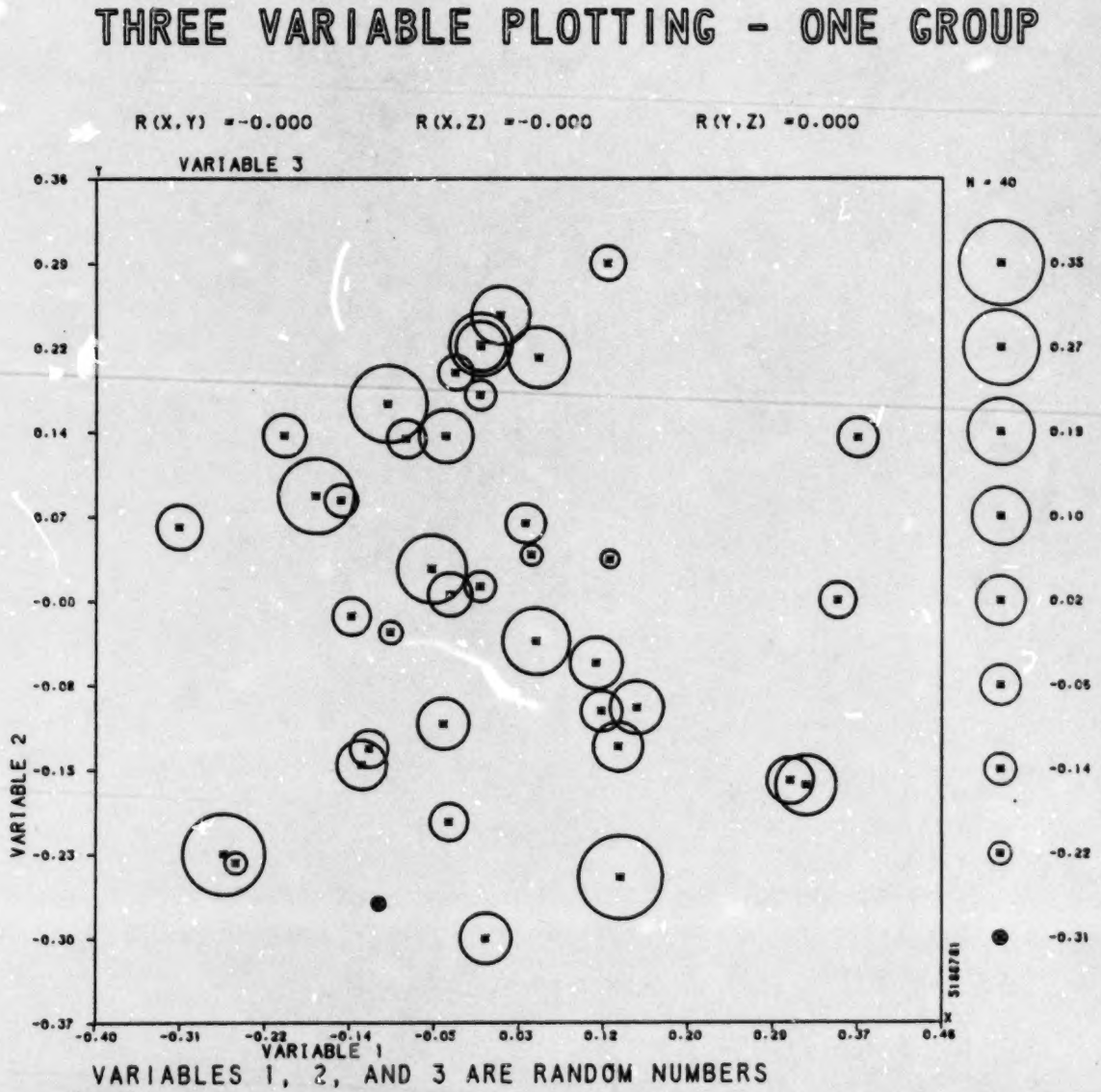

THREE VARIABLE PLOTTING - ONE GROUP
R(X,Y) =-0.000
R(X,Z) =-0.000
R(Y,Z) =0.000
VARIABLE 3
N = 40
0.35
0.27
0.19
0.10
0.02
-0.06
-0.14
-0.22
-0.31
0.36
0.29
0.22
0.14
0.07
-0.00
-0.08
-0.15
-0.23
-0.30
-0.37
VARIABLE 2
-0.40
-0.31
-0.22
-0.14
-0.05
0.03
0.12
0.20
0.29
0.37
0.46
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 4.

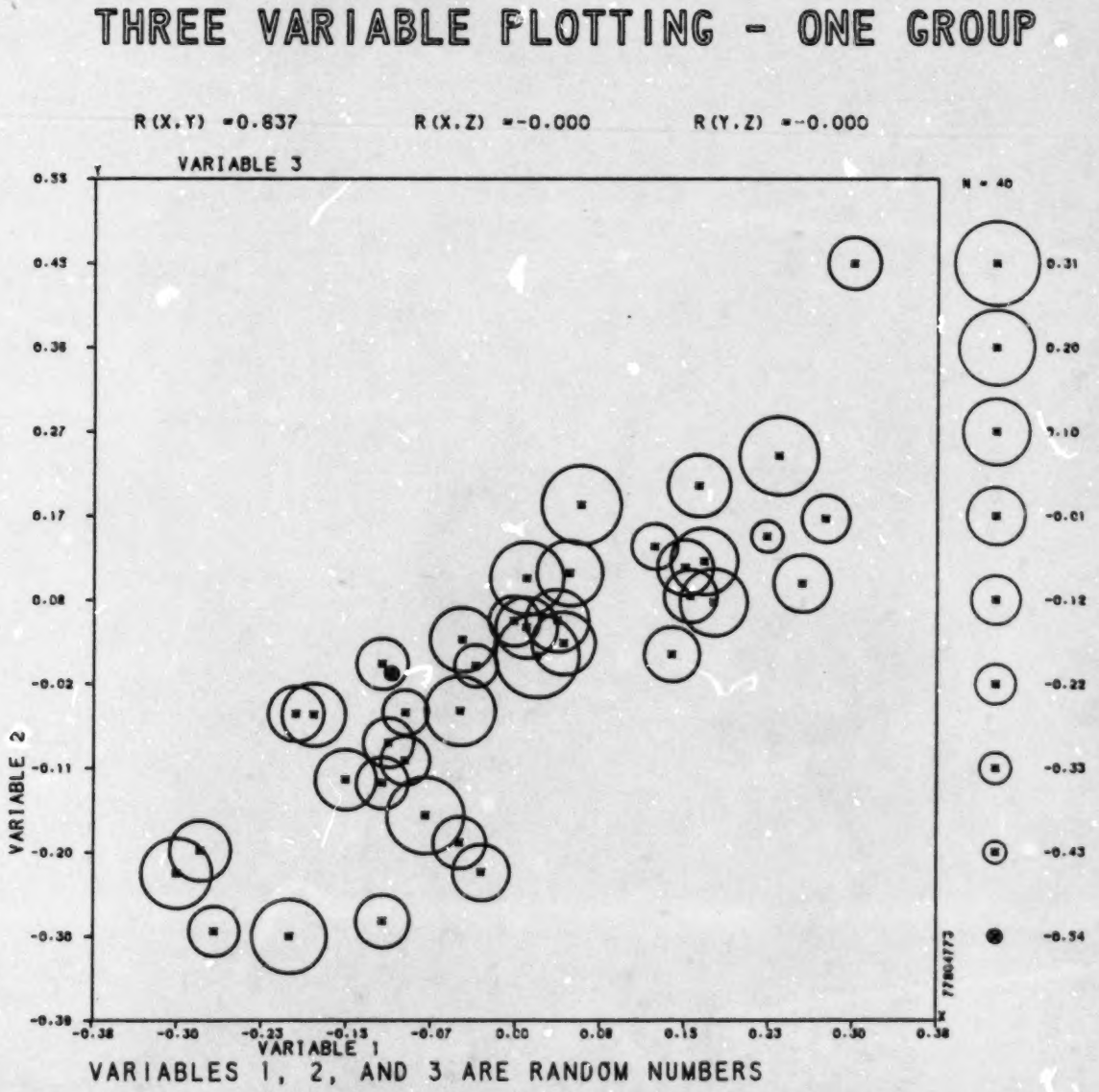
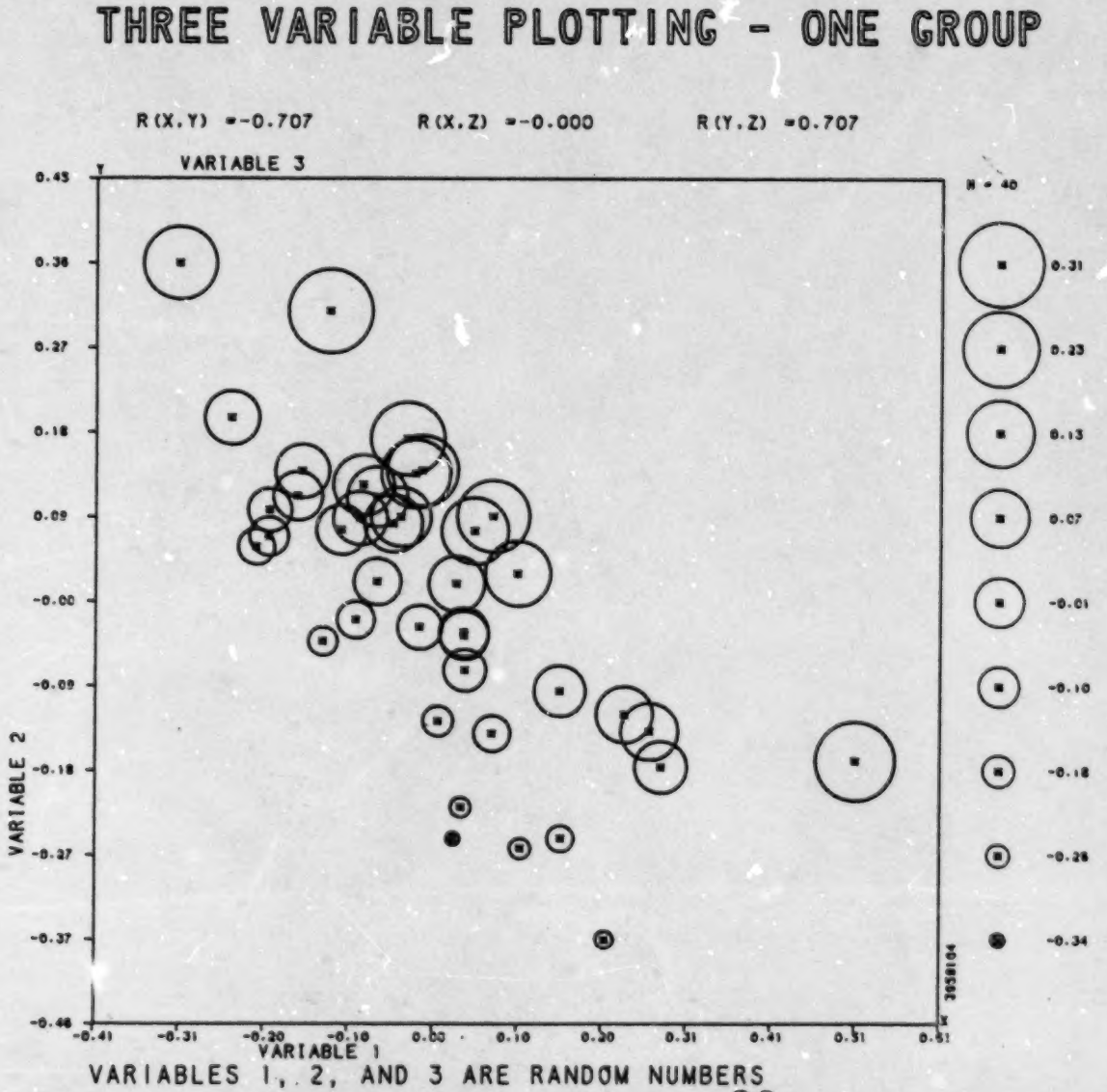

THREE VARIABLE PLOTTING - ONE GROUP
R(X,Y) =0.837
R(X,Z) =-0.000
R(Y,Z) =-0.000
VARIABLE 3
N = 40
0.31
0.20
0.10
-0.01
-0.12
-0.22
-0.33
-0.43
-0.54
0.53
0.45
0.36
0.27
0.17
0.08
-0.02
-0.11
-0.20
-0.30
-0.39
VARIABLE 2
-0.38
-0.30
-0.23
-0.15
-0.07
0.00
0.08
0.15
0.23
0.30
0.38
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 5.



VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 6.

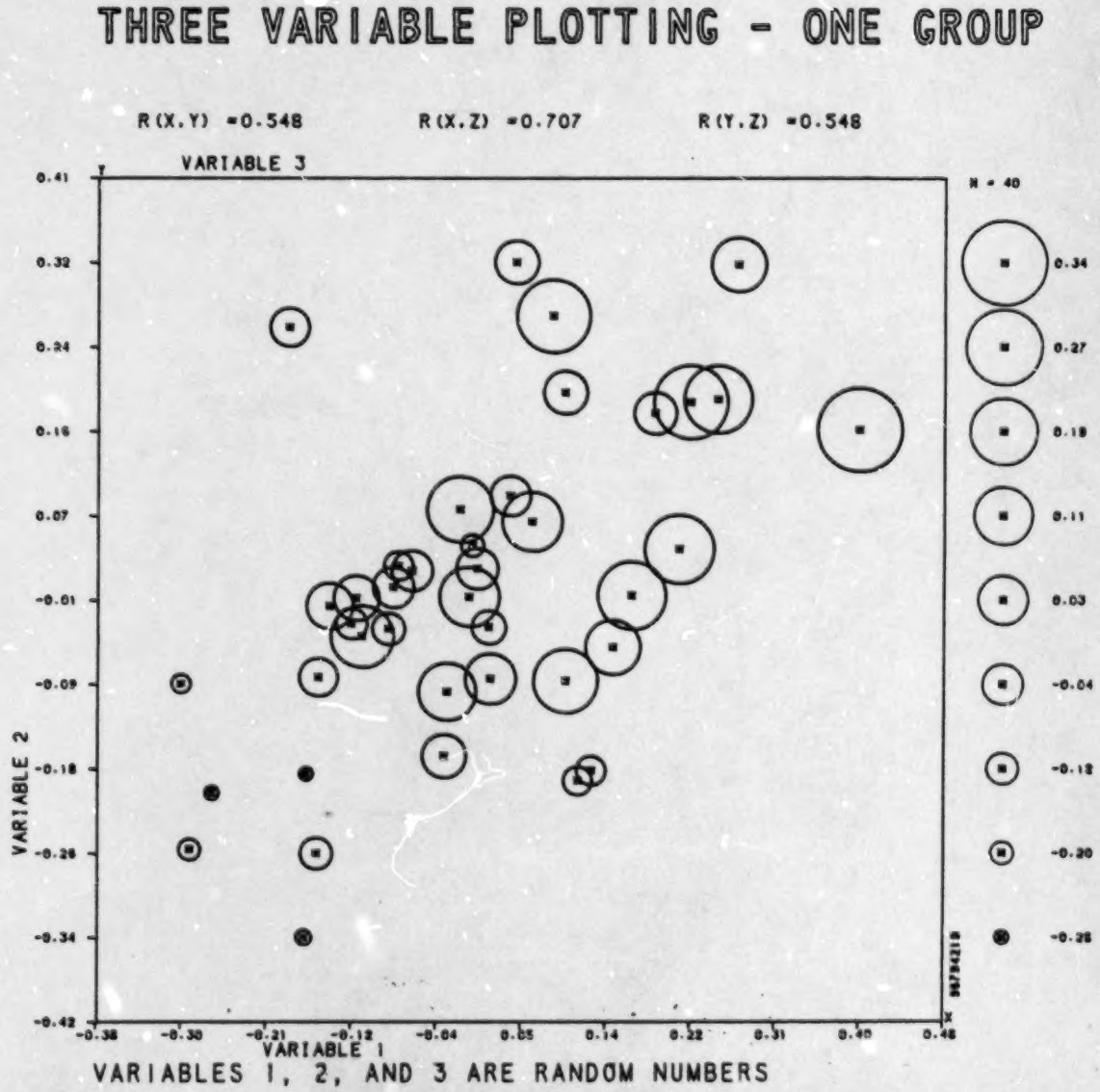

THREE VARIABLE PLOTTING - ONE GROUP
R(X,Y) =0.548
R(X,Z) =0.707
R(Y,Z) =0.548
VARIABLE 3
N = 40
0.41
0.32
0.24
0.16
0.07
-0.01
-0.09
-0.18
-0.26
-0.34
-0.42
0.34
0.27
0.19
0.11
0.03
-0.04
-0.12
-0.20
-0.28
VARIABLE 2
-0.38
-0.30
-0.21
-0.12
-0.04
0.05
0.14
0.22
0.31
0.40
0.48
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 7.

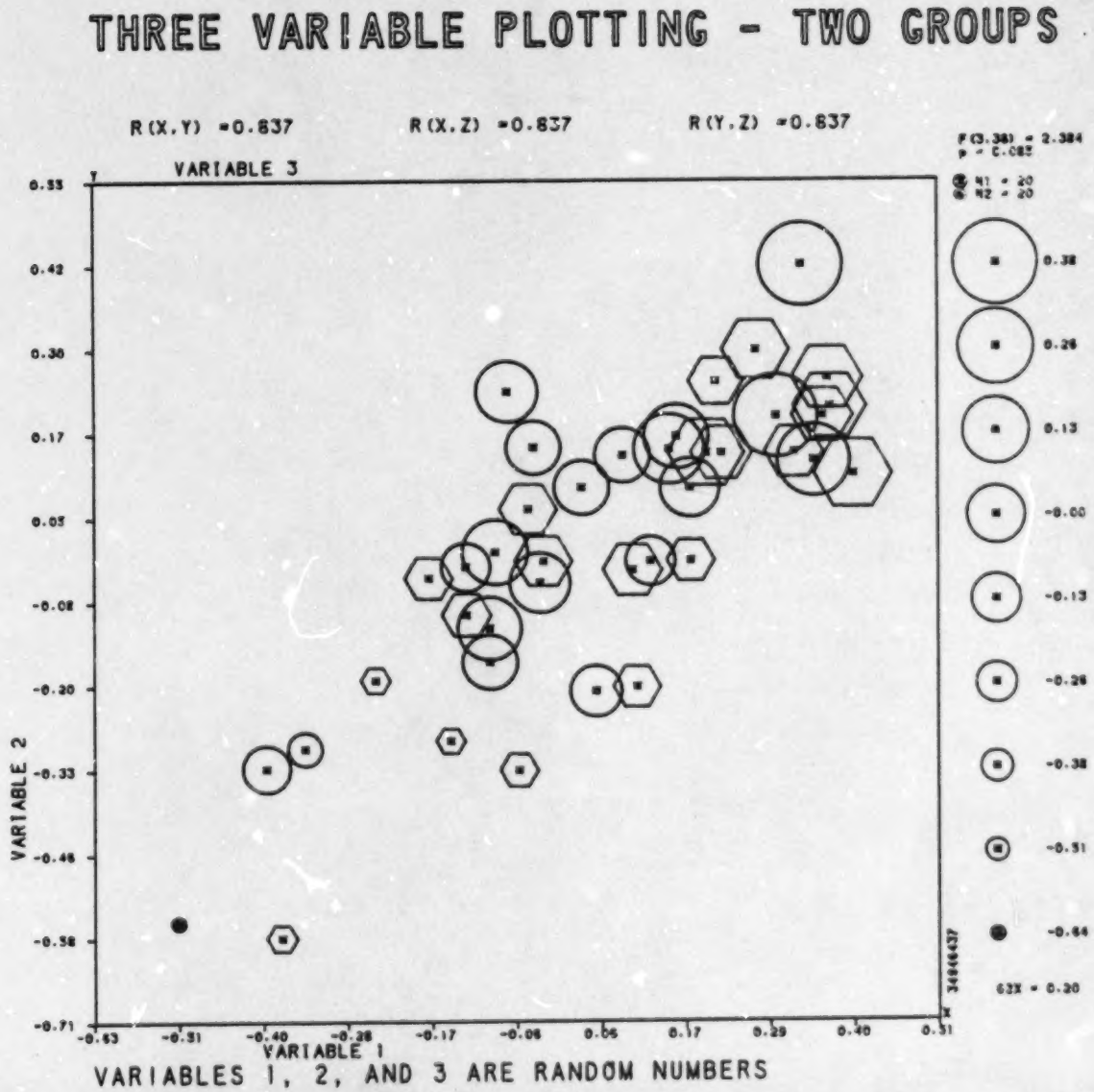

THREE VARIABLE PLOTTING - TWO GROUPS
R(X,Y) =0.837
R(X,Z) =0.837
R(Y,Z) =0.837
VARIABLE 3
VARIABLE 2
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 8.

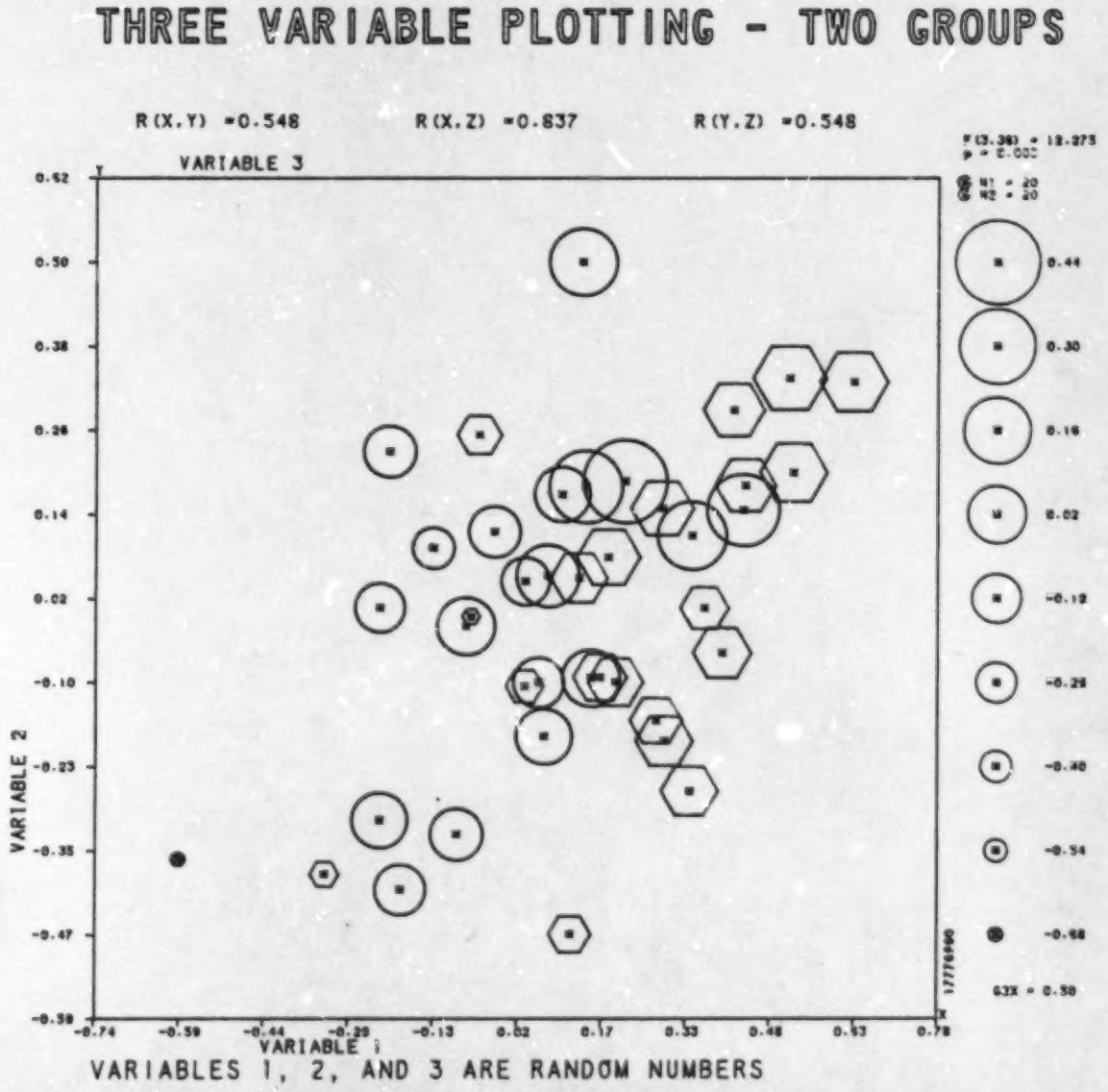

THREE VARIABLE PLOTTING - TWO GROUPS
R(X,Y) =0.548
R(X,Z) =0.837
R(Y,Z) =0.548
VARIABLE 3
VARIABLE 2
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 9.

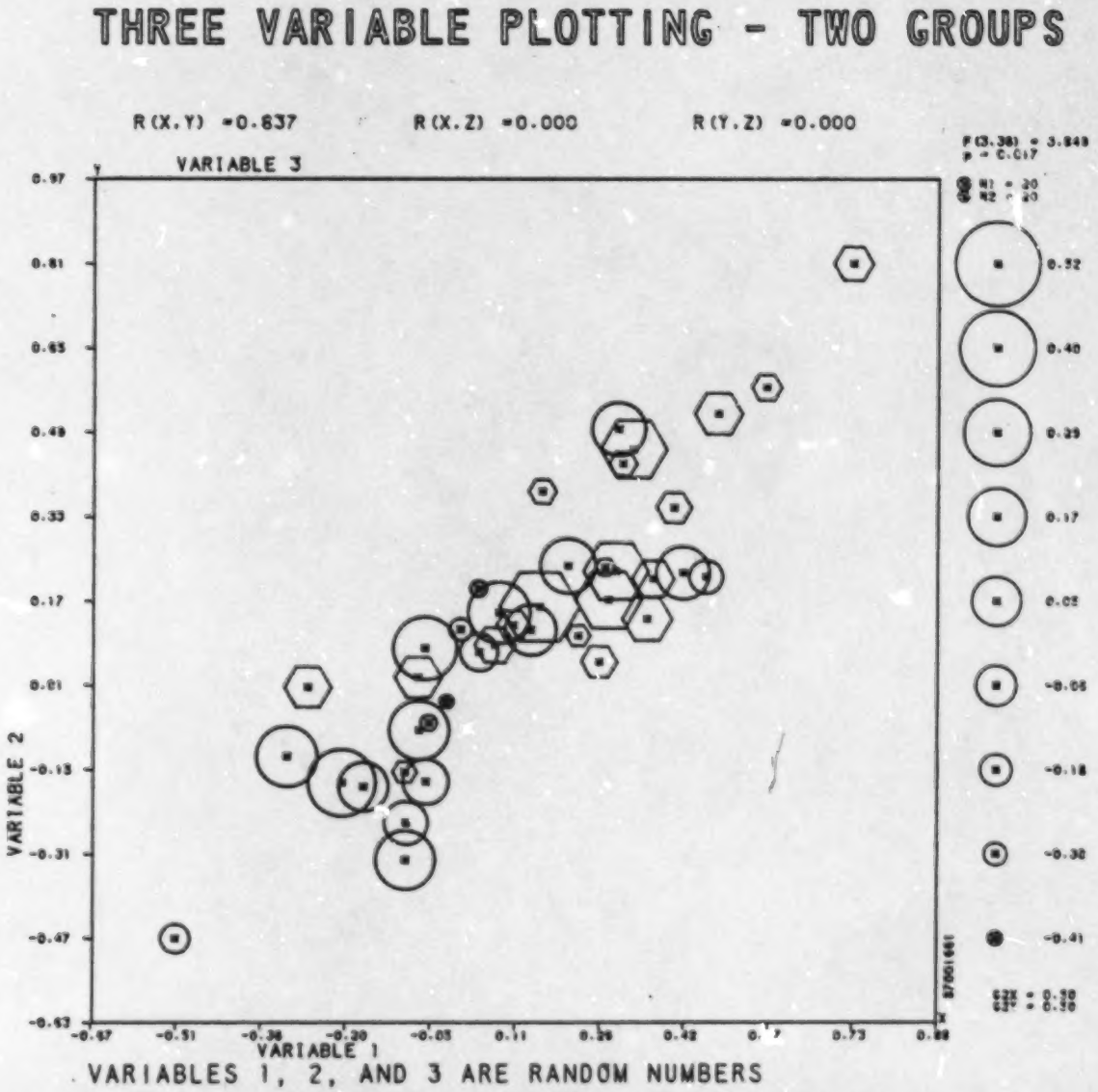

THREE VARIABLE PLOTTING - TWO GROUPS
R(X,Y) =0.837
R(X,Z) =0.000
R(Y,Z) =0.000
VARIABLE 3
VARIABLE 2
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 10.

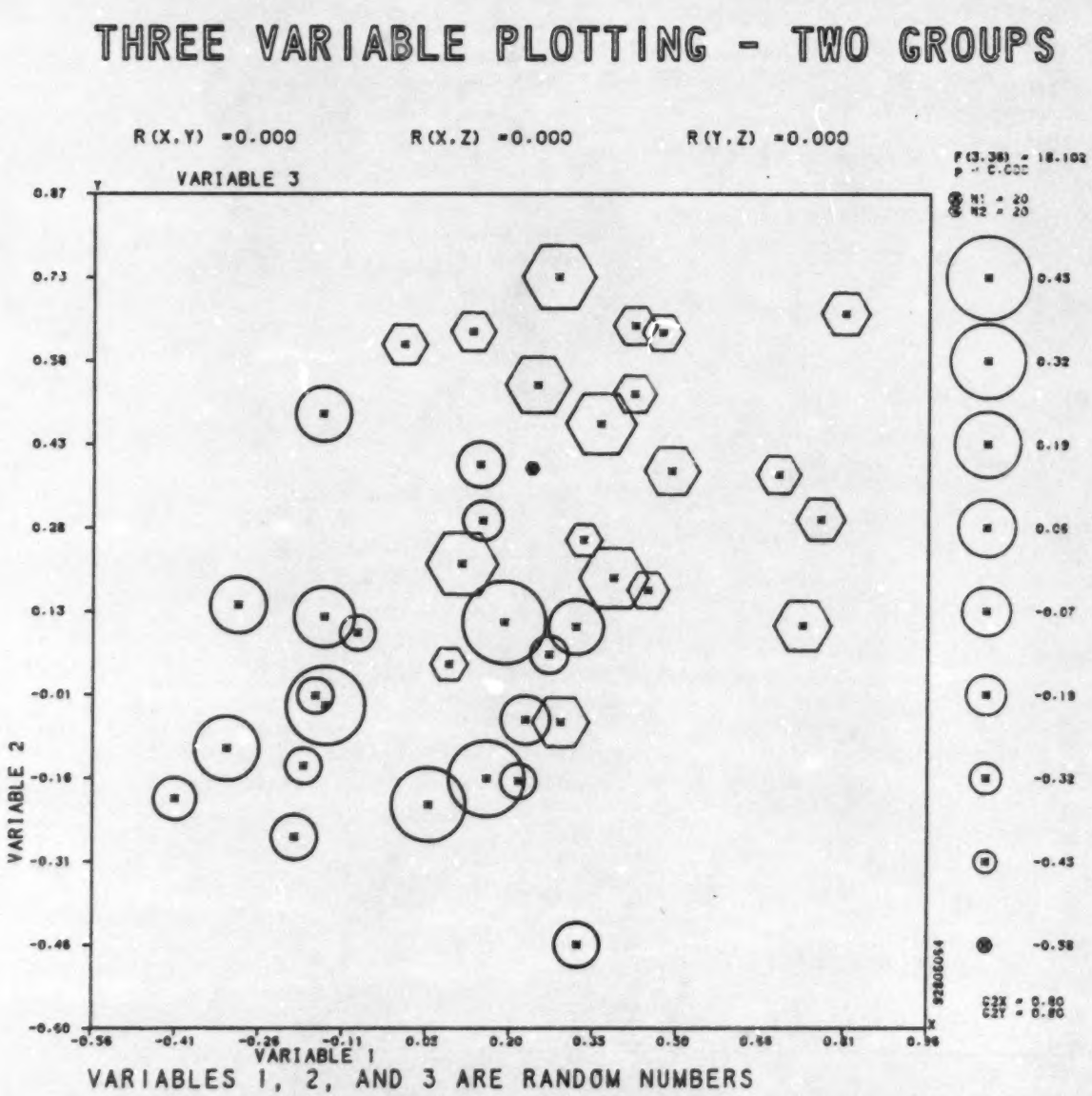

THREE VARIABLE PLOTTING - TWO GROUPS
R(X,Y) =0.000
R(X,Z) =0.000
R(Y,Z) =0.000
VARIABLE 3
VARIABLE 2
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 11.

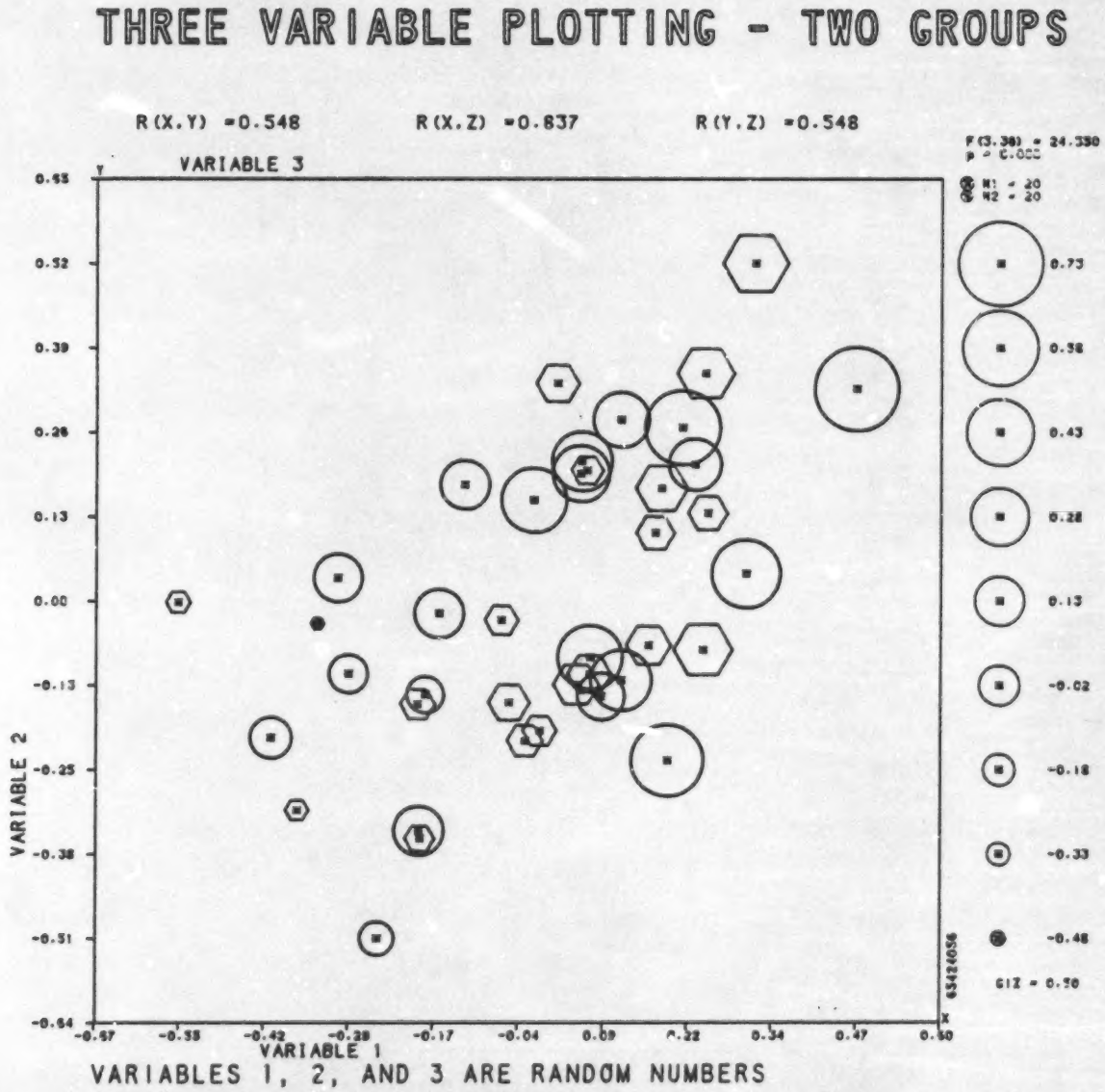
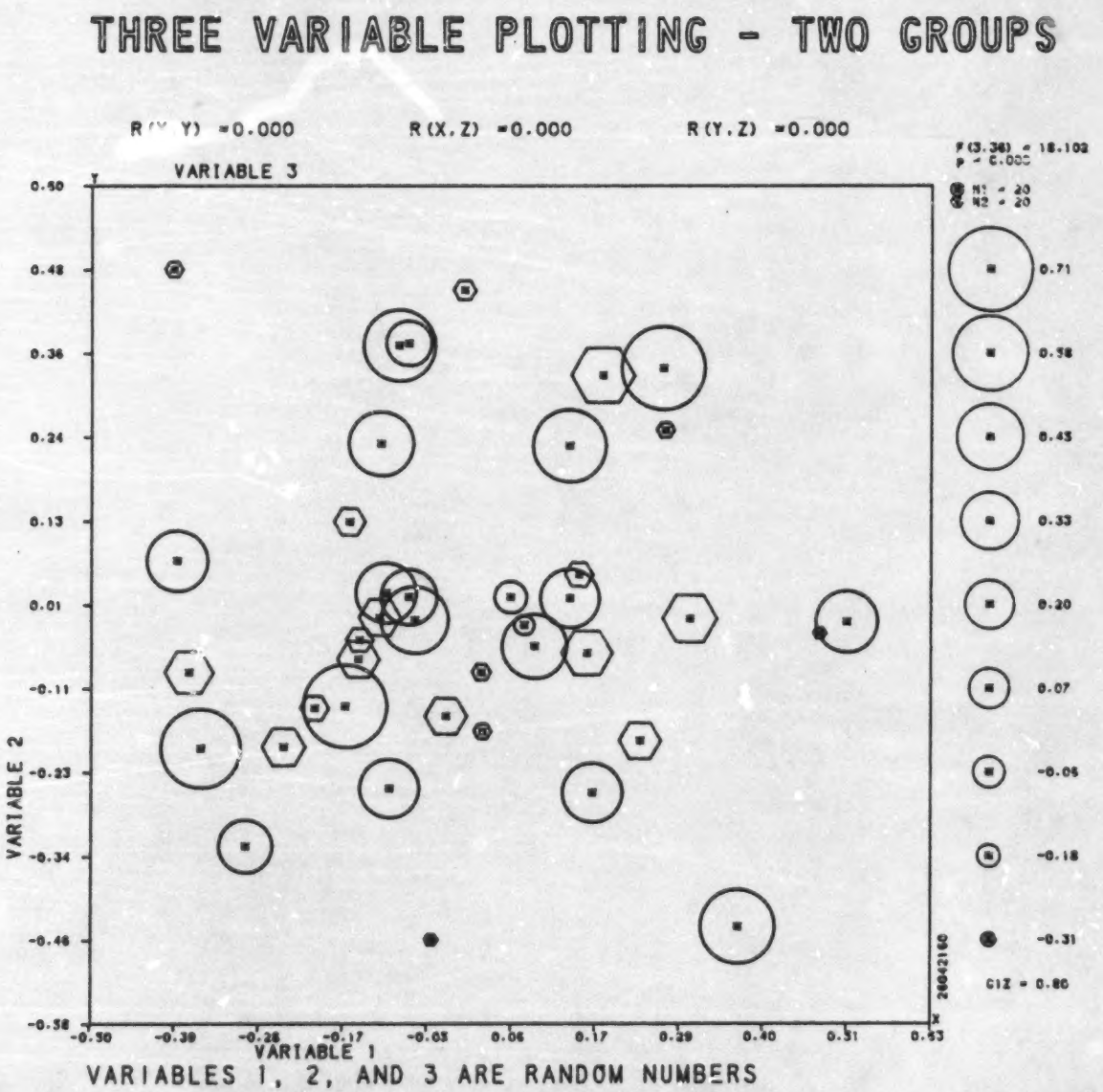

THREE VARIABLE PLOTTING - TWO GROUPS
R(X,Y) =0.548
R(X,Z) =0.837
R(Y,Z) =0.548
VARIABLE 3
VARIABLE 2
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 12.



VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 13.

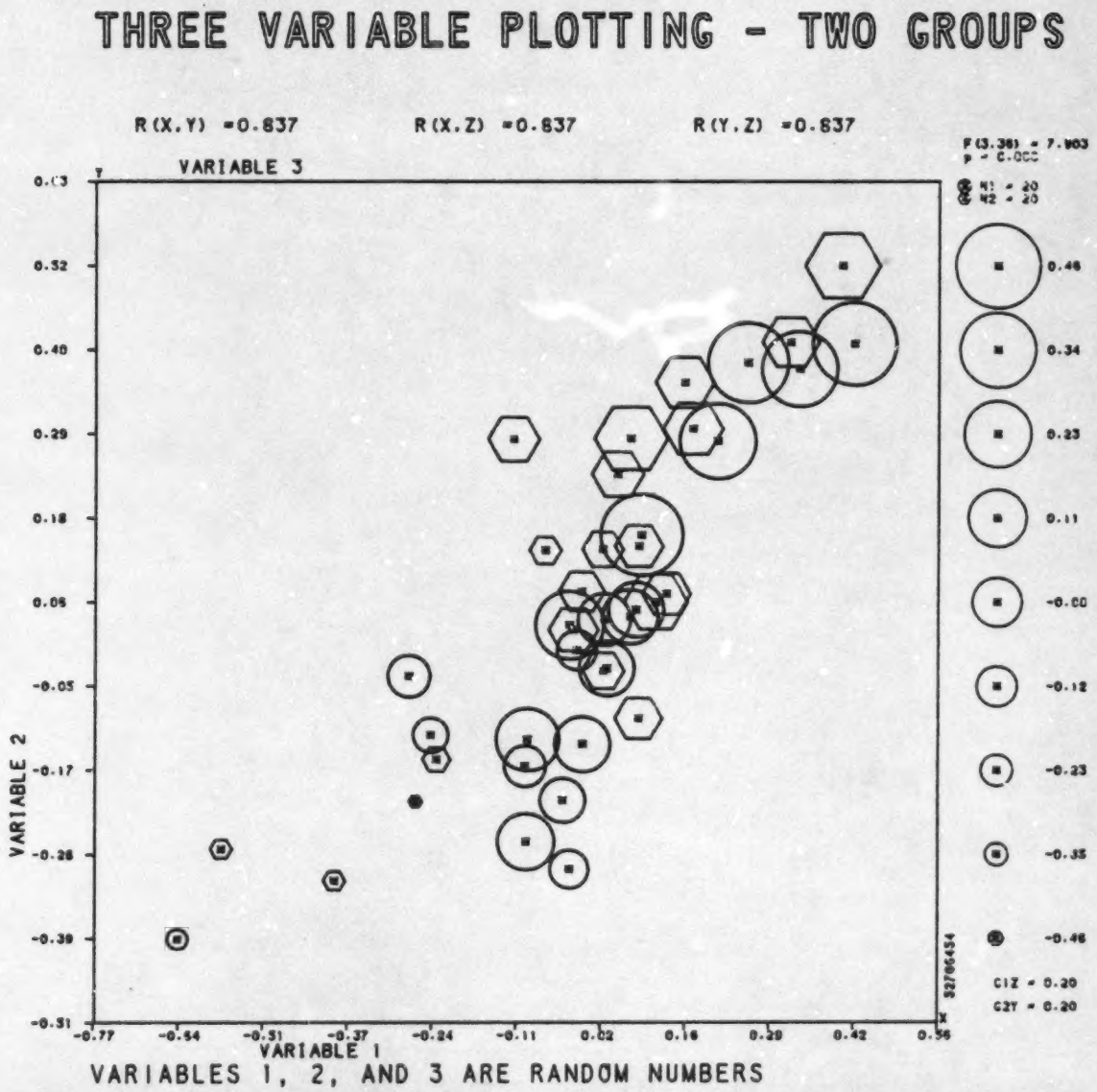

THREE VARIABLE PLOTTING - TWO GROUPS
R(X,Y) =0.837
R(X,Z) =0.837
R(Y,Z) =0.837
VARIABLE 3
VARIABLE 2
VARIABLE 1
0.46
0.34
0.23
0.11
-0.00
-0.12
-0.23
-0.35
-0.46
C1Z = 0.20
C2Y = 0.20
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 14.

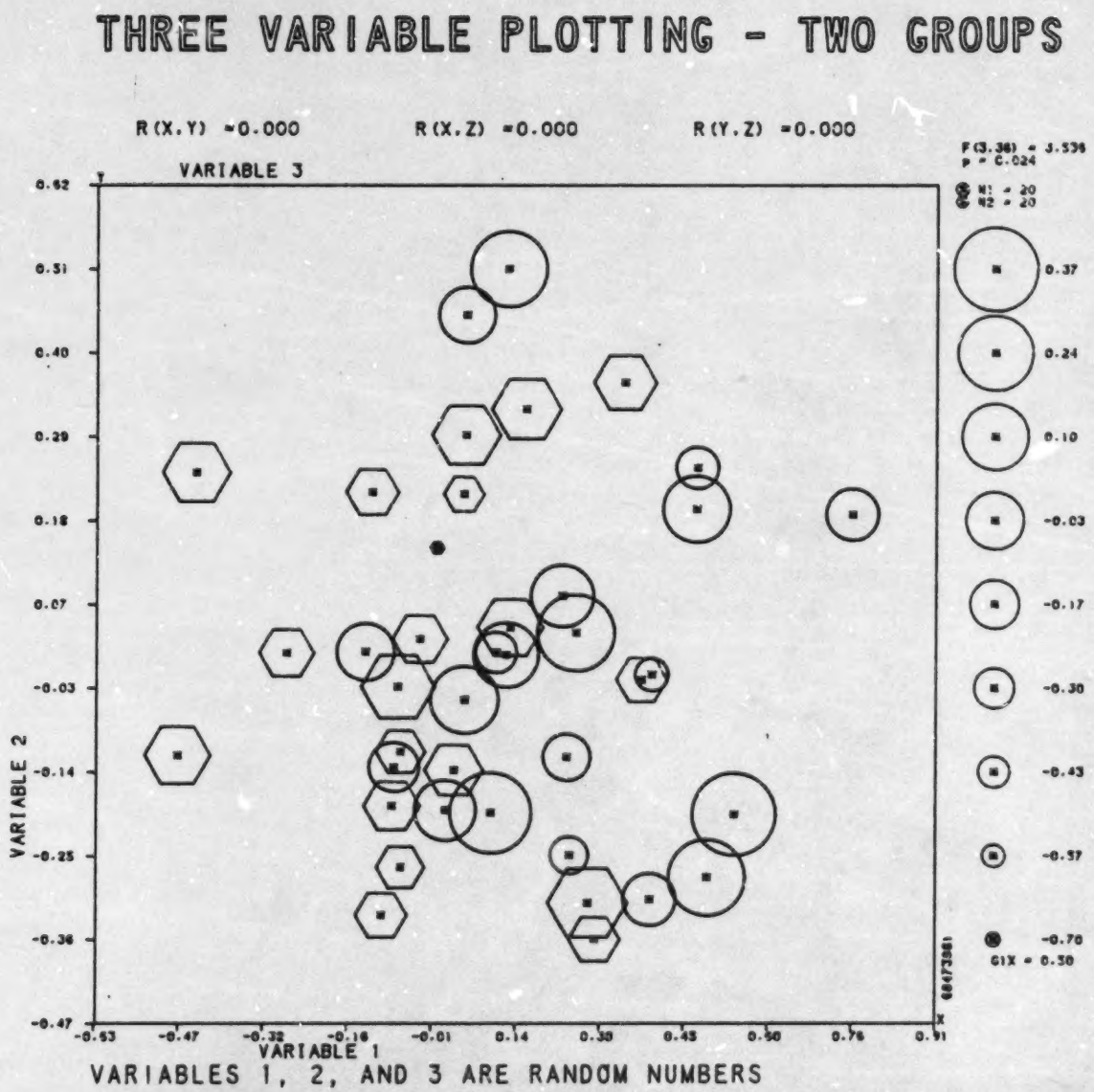
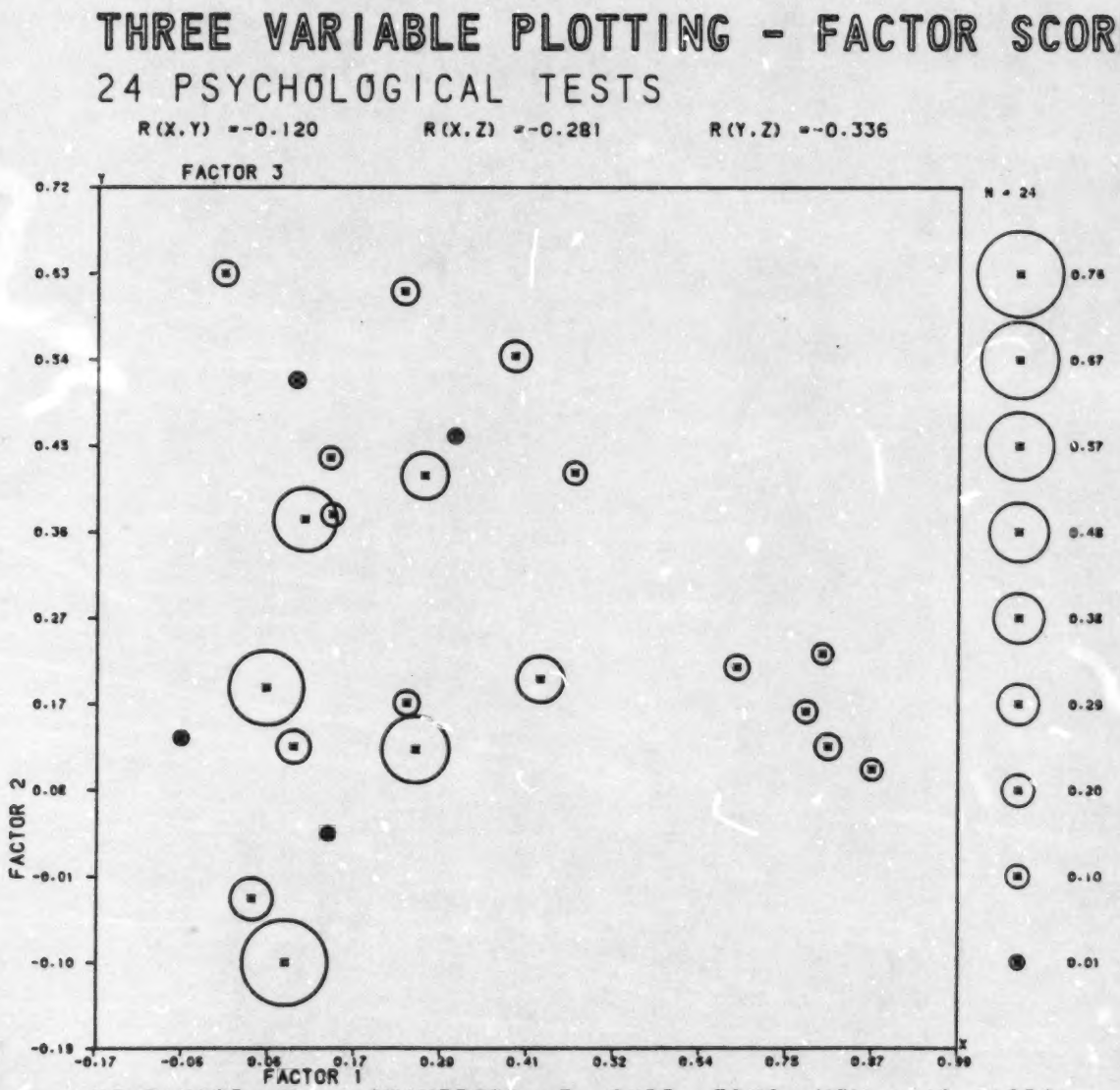
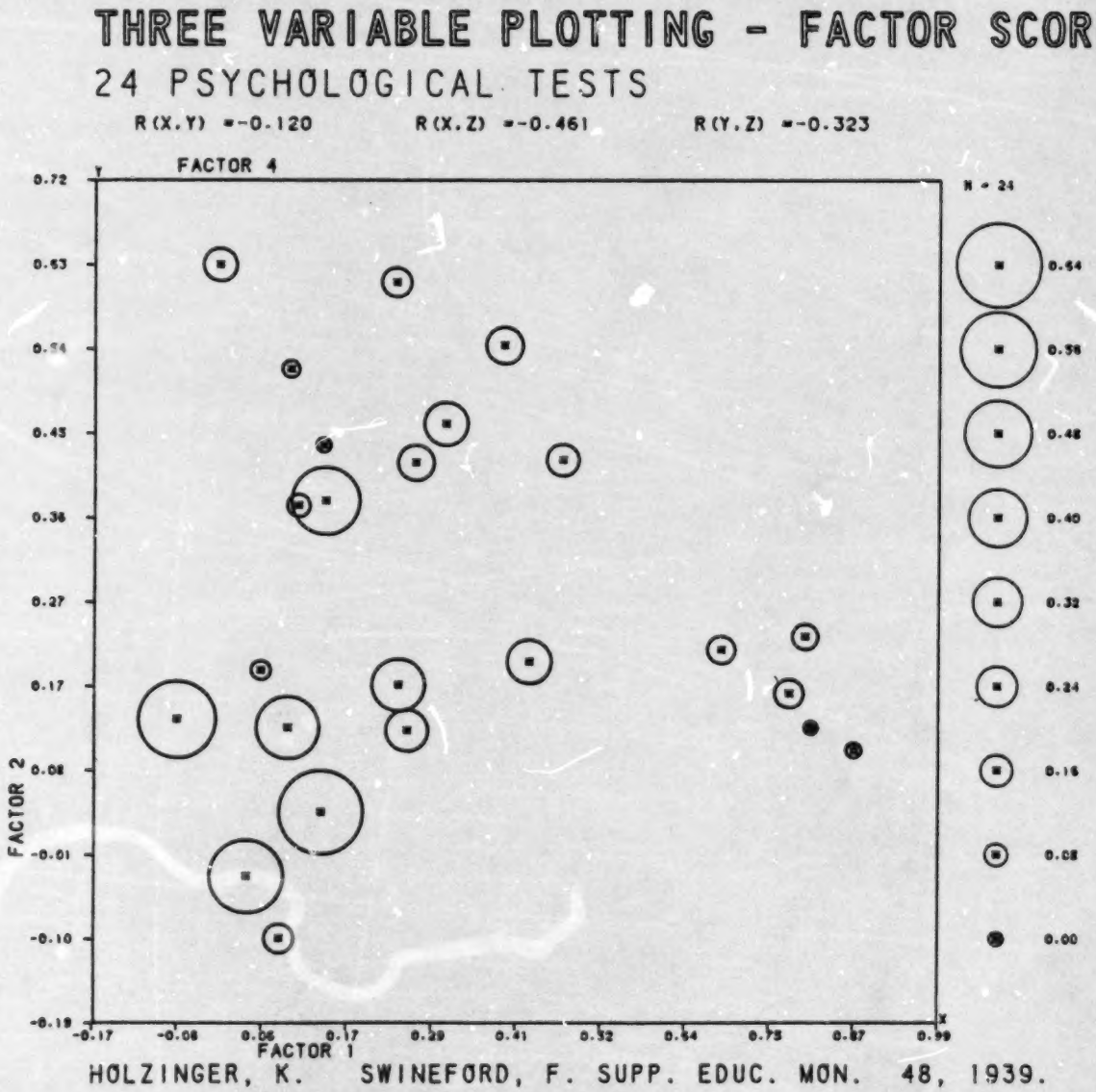

THREE VARIABLE PLOTTING - TWO GROUPS
R(X,Y) =0.000
R(X,Z) =0.000
R(Y,Z) =0.000
VARIABLE 3
VARIABLE 2
VARIABLE 1
VARIABLES 1, 2, AND 3 ARE RANDOM NUMBERS

Figure 15.



HOLZINGER, K. SWINEFORD, F. SUPP. EDUC. MON. 48, 1939.

Figure 16.



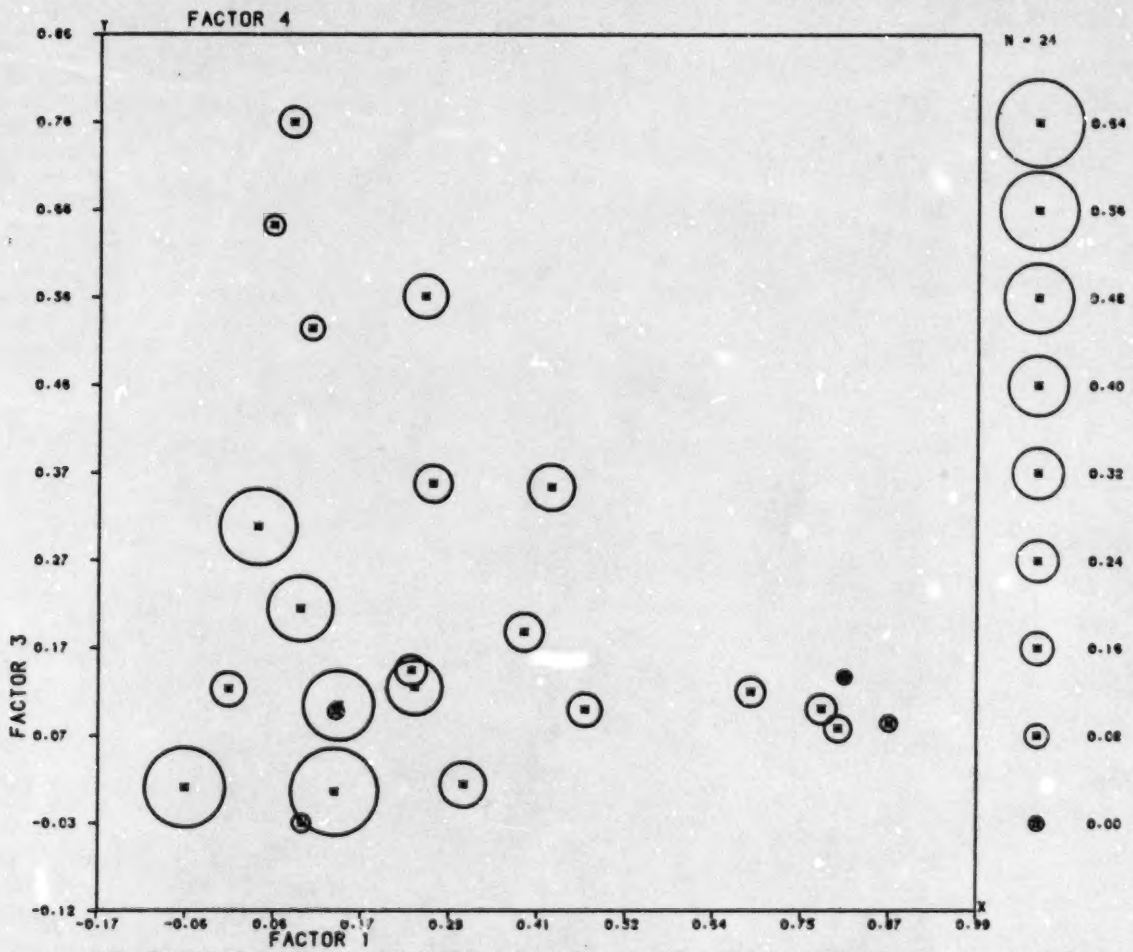
HOLZINGER, K. SWINEFORD, F. SUPP. EDUC. MON. 48, 1939.

Figure 17.

THREE VARIABLE PLOTTING - FACTOR SCORES

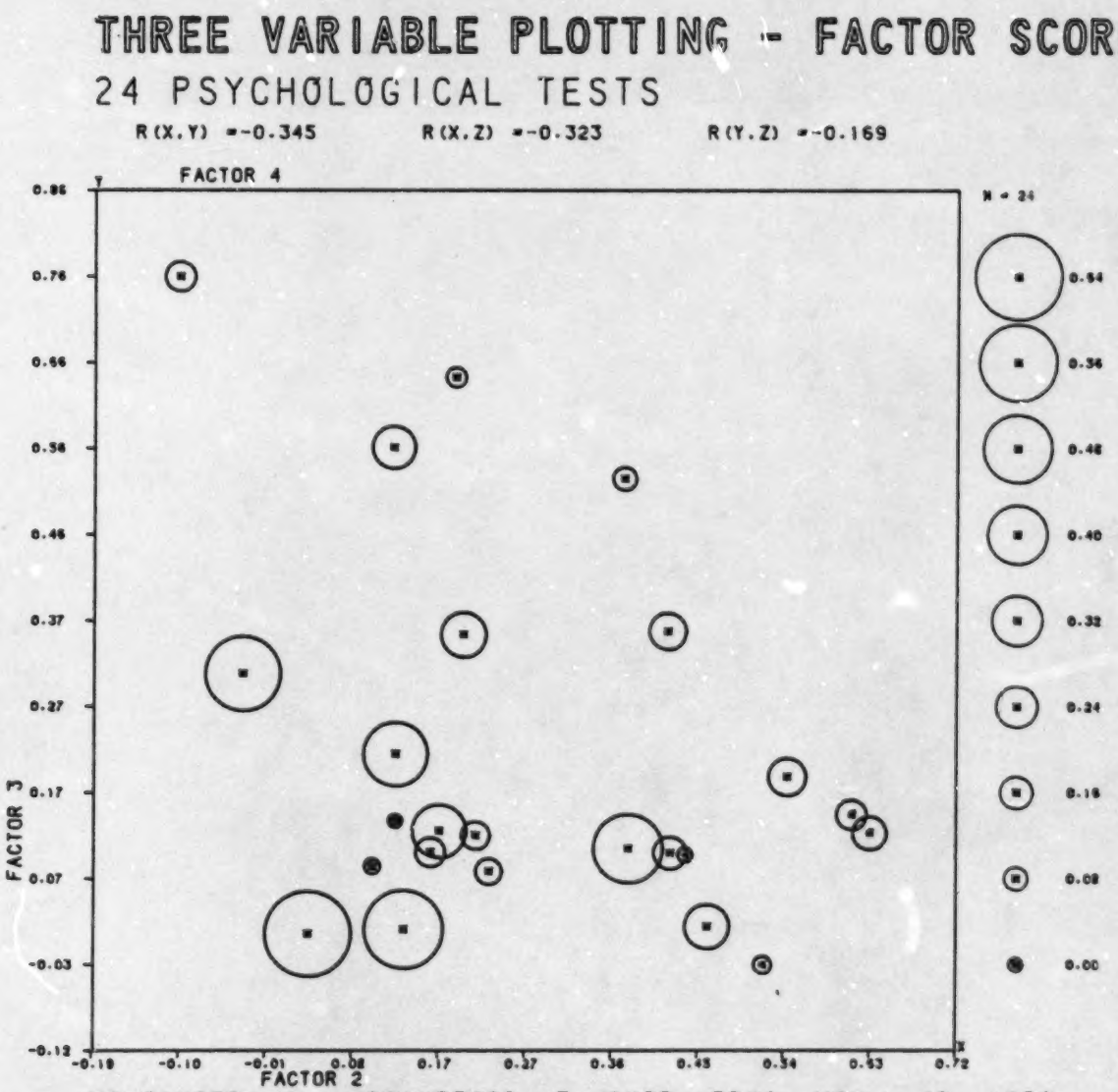
2[illegible] PSYCHOLOGICAL TESTS

R(X,Y) =-0.271 R(X,Z) =-0.461 R(Y,Z) =-C 169



HOLZINGER, K. SWINEFORD, F. SUPP. EDUC. MON. 48, 1939.

Figure 18.



HOLZINGER, K. SWINEFORD, F. SUPP. EDUC. MON. 48, 1939.